(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
18 October 2001 (18.10.2001)

PCT

(10) International Publication Number
WO 2001/077900 A3

(51) International Patent Classification[7]: G06F 17/30

(21) International Application Number: PCT/US2001/011325

(22) International Filing Date: 5 April 2001 (05.04.2001)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
60/195,556 6 April 2000 (06.04.2000) US
09/728,689 1 December 2000 (01.12.2000) US

(71) Applicant: ISPHERES CORPORATION [US/US]; 640 Third Street, Oakland, CA 94607 (US).

(72) Inventors: BAX, Eric, T.; 1267 N. Michigan Avenue, Pasadena, CA 91104 (US). FOWLKES, Charless, C.; 31 Gardner Park Drive, Bozeman, MT 59715 (US). CISNERO, Louis, Jr.; 1208 Commerce Street, Jourdanton, TX 78026 (US).

(74) Agent: MCKENZIE, Denise, L.; Sidley Austin Brown & Wood, 555 West Fifth Street, Los Angeles, CA 90013-1010 (US).

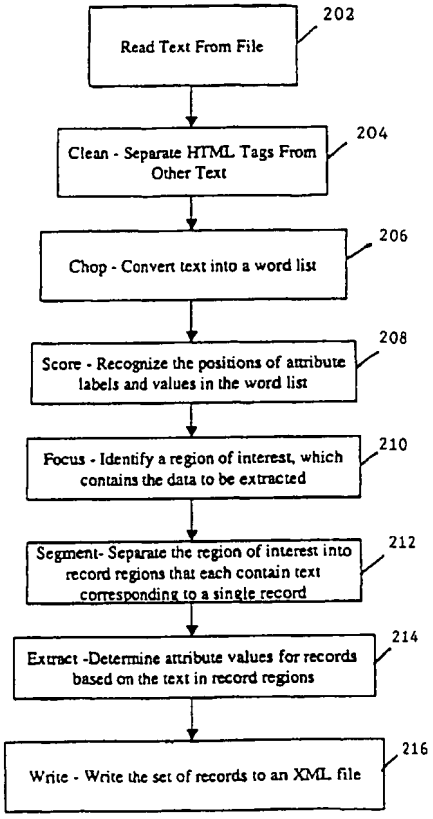
(81) Designated States *(national)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CR, CU, CZ, DE, DK, DM, DZ, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, TZ, UA, UG, UZ, VN, YU, ZA, ZW.

(84) Designated States *(regional)*: ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).

Published:
— *with international search report*

[Continued on next page]

(54) Title: TECHNIQUE FOR EXTRACTING DATA FROM STRUCTURED DOCUMENTS



(57) Abstract: The present invention discloses a technique for extracting data from a file. In accordance with the present invention, a request to extract one or more data records from the file is received. The data records within the file are identified, without using prior knowledge of a structure of the file. The data records are then extracted.

WO 2001/077900 A3

— *before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments*

(88) Date of publication of the international search report:
1 April 2004

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

INTERNATIONAL SEARCH REPORT

International Application No
PCT/US 01/11325

A. CLASSIFICATION OF SUBJECT MATTER
IPC 7 G06F17/30

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
IPC 7 G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)
WPI Data, PAJ, INSPEC, EPO-Internal

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category *	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	HAMMER J ET AL: "Extracting semistructured information from the Web" PROCEEDINGS OF THE WORKSHOP ON MANAGEMENT OF SEMI-STRUCTURED DATA, XX, XX, 16 March 1997 (1997-03-16), pages 1-8-25, XP002103690 the whole document ---	1-42
A	KUSHMERICK N ET AL: "WRAPPER INDUCTION FOR INFORMATION EXTRACTION" PROCEEDINGS OF THE INTERNATIONAL JOINT CONFERENCE ON ARTIFICIAL INTELLIGENCE, XX, XX, no. 1, 23 August 1997 (1997-08-23), pages 729-735, XP001079939 paragraphs 1-4 ---	1-42

-/--

[X] Further documents are listed in the continuation of box C.

[ ] Patent family members are listed in annex.

* Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
12 January 2004	18/02/2004

Name and mailing address of the ISA	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Tx. 31 651 epo nl, Fax: (+31-70) 340-3016	Perez Perez, J

Form PCT/ISA/210 (second sheet) (July 1992)

INTERNATIONAL SEARCH REPORT

International Application No
PCT/US 01/11325

C.(Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT

Category °	Citation of document, with indication,where appropriate, of the relevant passages	Relevant to claim No.
A	ADELBERG B: "NODOSE - A TOOL FOR SEMI-AUTOMATICALLY EXTRACTING STRUCTURED AND SEMISTRUCTURED DATA FROM TEXT DOCUMENTS" SIGMOD RECORD, SIGMOD, NEW YORK, NY, US, vol. 27, June 1998 (1998-06), pages 283-294, XP001080524 ISSN: 0163-5808 paragraphs 1,2 ---	1-42
A	S. ST. LAURENT: "Schematron: an interview with Rick Jelliffe" XMLHACK WEBSITE, 'Online! 15 November 1999 (1999-11-15), XP002261875 Retrieved from the Internet: <URL:http://www.xmlhack.com/read.php?item=121> 'retrieved on 2003-11-17! the whole document -----	1-42